Language Learning & Technology
http://llt.msu.edu/vol1num1/brown/default.html

July 1997, Volume 1, Number 1
pp.44-59 online
(page numbers in PDF differ and should not be used for reference)

COMPUTERS IN LANGUAGE TESTING: PRESENT RESEARCH AND SOME FUTURE DIRECTIONS

James Dean Brown
University of Hawai'i at Manoa

ABSTRACT

This article begins by exploring recent developments in the use of computers in language testing in four areas: (a) item banking, (b) computer-assisted language testing, (c) computerized-adaptive language testing, and (d) research on the effectiveness of computers in language testing.

The article then examines the educational measurement literature in an attempt to forecast the directions future research on computers in language testing might take and suggests addressing the following issues: (a) piloting practices in computer adaptive language tests (CALTs), (b) standardizing or varying CALT lengths, (c) sampling CALT items, (d) changing the difficulty of CALT items, (e) dealing with CALT item sets, (f) scoring CALTS, (g) dealing with CALT item omissions, (h) making decisions about CALT cut-points, (i) avoiding CALT item exposure, (j) providing CALT item review opportunities, and (k) complying with legal disclosure laws when using CALTs.

The literature on computer-assisted language learning indicates that language learners have generally positive attitudes toward using computers in the classroom (Reid, 1986; Neu & Scarcella, 1991; Phinney, 1991), and a fairly large literature has developed examining the effectiveness of computer-assisted language learning (for a review, see Dunkel, 1991). But less is known about the more specific area of computers in language testing. The purpose of this article is to examine recent developments in language testing that directly involve computer use including what we have learned in the process. The article will also examine the dominant issue of computer-adaptive testing in the educational measurement literature in an attempt to forecast some of the directions future research on computers in language testing might take.

CURRENT STATE OF KNOWLEDGE ON COMPUTERS IN LANGUAGE TESTING

In reviewing the literature on computers in language testing, I have found four recurring sets of issues: (a) item banking, (b) computer-assisted language testing, (c) computer-adaptive language testing, and (d) the effectiveness of computers in language testing. The discussion in this section will be organized under those four headings.

Item Banking

Item banking covers any procedures that are used to create, pilot, analyze, store, manage, and select test items so that multiple test forms can be created from subsets of the total "bank" of items. With a large item bank available, new forms of tests can be created whenever they are needed. Henning (1986) provides a description of how item banking was set up for the ESL Placement Examination at UCLA. (For further explanation and examples of item banking in educational testing, see Baker, 1989, pp. 412-414, or Flaugher, 1990.)

While the underlying aims of item banking can be accomplished by using traditional item analysis procedures (usually item facility and item discrimination indexes; for a detailed description of these

traditional item analysis procedures, see Brown, 1996), a problem often occurs because of differences in abilities among the groups of people who are used in piloting the items, especially when they are compared to the population of students with whom the test is ultimately to be used. However, a relatively new branch of test analysis theory, called item response theory (IRT), eliminates the need to have exactly equivalent groups of students when piloting items because IRT analysis yields estimates of item difficulty and item discrimination that are "sample-free." IRT can also provide "item-free" estimates of students' abilities. Naturally, a full discussion of IRT is beyond the scope of this article. However, Henning (1987) discusses the topic in terms of the steps involved in item banking for language tests and provides recipe-style descriptions of how to calculate the appropriate IRT statistics.

Several other references may prove helpful for readers interested in more information on IRT. In language testing, Madsen and Larson (1986) use computers and IRT to study item bias, while de Jong (1986) demonstrates the use of IRT for item selection purposes. (For readers who want more technical information on applications of IRT to practical testing problems in general education, see Lord, 1980; Hambleton & Swaminathan, 1985; Andrich, 1988; Suen, 1990; Wainer & Mislevy, 1990; and Hambleton, Swaminathan, & Rogers, 1991.)

I am not saying that item banking is without potential problems. Green (1988) outlines some of the problems that might be encountered in using IRT in general, and Henning (1991) discusses specific problems that may be encountered with the validity of item banking techniques in language testing settings. Another serious limitation of IRT is the large number of students that must be tested before it can responsibly be applied. Typically, IRT is only applicable for full item analysis (that is, for analysis of two or three parameters) when the numbers of students being tested are very large by the standards of most language programs, that is to say, in excess of one thousand. Smaller samples in the hundreds can be used only if the item difficulty parameter is studied.

Minimal item banking can be done without computers by using file cards, and, of course, the traditional item analysis statistics can be done (using the sizes of groups typically found in language programs) with no more sophisticated equipment than a hand-held calculator. Naturally, a personal computer can make both item banking and item analysis procedures much easier and much faster. For example, standard database software can be used to do the item banking, (e.g., Microsoft Access, 1996; or Corel Paradox, 1996). For IRT analyses, more specialized software will be needed. The following are examples of computer programs that can be used for IRT analysis: TESTAT (Stenson, 1988), BIGSTEPS (Wright, Linacre, & Schulz, 1990), and PC-BILOG (Mislevy & Bock, 1986). Alternatively, the MicroCAT Testing System (1984) program can help with both item banking and IRT analyses. BestTest (1990) is another less sophisticated program, which can be used in both item banking and test creation. An example of a software program specifically designed for item banking is the PARTest (1990) program. If PARTest is used in conjunction with PARScore (1990) and PARGrade (1990), a completely integrated item banking, test analysis, and record-keeping system can be set up and integrated with a machine scoring system. Indications have also surfaced that computers may effectively be used to assemble pre-equated language tests from a bank of items (see Henning, Johnson, Boutin, & Rice, 1994).

Computer-Assisted Language Testing

Tests that are administered at computer terminals, or on personal computers, are called computer-assisted tests. Receptive-response items-including multiple-choice, true-false, and matching items-are fairly easy to adapt to the computer-assisted testing medium. Relatively cheap authoring software like *Testmaster* (1988) can be used to create such tests. Even productive-response item types-including fill-in and cloze-can be created using authoring software like Testmaster. Unfortunately, the more

interesting types of language tasks (e.g., role plays, interviews, compositions, oral presentations) prove much more difficult to develop for computer-assisted testing.

However, advancing technologies have many potential ramifications for computer-assisted language testing. Brown (1992a) outlined some of the technological advances that may have an impact on language teaching and testing:

> Consider the multi-media combinations that will be available in the very near future: CD-ROM players working with video-image projectors, and computers controlling the whole interactive process between students and machines for situations and language tailored to each student's specific needs....Consider the uses to which computer communications networks could be put. What about scanners and hand-writing recognition devices? Won't voice sensitive computers and talking computers be valuable tools in the language media services of the future? (p. 2)

The new technologies such as the CD-ROM and interactive video discussed in Brown (1992a) do make it possible for students to interact with a computer. Hence, no technical reason remains why interactive testing like role plays, interviews, compositions, and presentations cannot be done in a computer-assisted mode. Naturally, the expense involved may impose some limits, and the scoring will probably continue to involve rater judgments (thus, further increasing the expense involved). But at least, the logistics of gathering the language samples can now be simplified by the use of computer-assisted testing procedures.

Two consequences may evolve from the current advances in technology: (a) the sophistication of existing computer hardware and software tools will continue to grow, and (b) the cost of the technology will continue to drop (eventually to within reach of all language programs). Hence, the possibilities for developing productive-response computer-assisted language tests will definitely increase.

But, why should we bother to create computer-assisted language tests at all? Aren't they really just a sophisticated version of the paper-and-pencil tests that they will probably be modeled on? Two primary benefits can be gained from computer-assisted language testing:

1. Computer-assisted language tests can be individually administered, even on a walk-in basis. Thus group-administered tests and all of the organizational constraints that they impose will no longer be necessary.
2. Traditional time limits are not necessary. Students can be given as much time as they need to finish a given test because no human proctor needs to wait around for them to finish the test.

No doubt, cheating will arise, but such problems can be surmounted if a little thought and planning are used.

Given the advantages of individual, time-independent language testing, computer-assisted testing will no doubt prove to be a positive development. Consider the benefits of a writing test administered in a computer laboratory as the final examination for an ESL writing course. Such a computer-assisted test would be especially suitable for students who had been required to do all of their writing assignments in the course on a PC. In such a course, it would make imminent sense to allow the students to do their final examination writing samples on a computer and turn in the diskette at the end of the testing period (or send the file by modem or network to the teacher). Under such circumstances, the testing period could be quite long to allow time for multiple revisions. Of course, logistical problems will crop up, but they can no doubt be overcome with careful planning. In fact, the literature indicates that computers can be an effective tool for teaching writing (Neu & Scarcella, 1991; Phinney, 1991). Why not also use it as an effective tool for testing writing (see for example Reid, 1986)?

Computer-Adaptive Language Testing

Computer-adaptive language tests are a subtype of computer-assisted language tests because they are administered at computer terminals or on personal computers. The computer-adaptive subtype of computer-assisted tests has three additional characteristics: (a) the test items are selected and fitted to the individual students involved, (b) the test is ended when the student's ability level is located, and, as a consequence, (c) computer-adaptive tests are usually relatively short in terms of the number of items involved and the time needed. As Madsen (1991) put it, "The computer-adaptive language test (CALT) is uniquely tailored to each individual. In addition, CALT is automatically terminated when the examinee's ability level has been determined....The result is a test that is more precise yet generally much shorter than conventional paper-and-pencil tests" (p. 237).

A clear description of how to develop computer-adaptive language tests (CALTs) is provided in Tung (1986). (For descriptions of more general computer-adaptive test [CAT] development in educational measurement, see Kaya-Carton, Carton, & Dandonoli, 1991, as well as Laurier, 1991; 1996.) CALT development relies very much on item response theory. While the computer-adaptive language test is taking place, the computer typically uses a combination of item response theory and the concept of flexilevel tests (Lord, 1980) to create a test specifically designed for the individual student taking it.

The flexilevel procedures roughly determine the general ability level of the student within the first few test questions. Then, based on item response statistics, the computer selects items which are suitable for the student's particular level and administers those items in order to get a more finely tuned estimate of the student's ability level. This flexilevel strategy eliminates the need (usually present in traditional fixed-length paper-and-pencil tests) for students to answer numerous questions that are too difficult or too easy for them. In fact, in a CALT, all students take tests that are suitable to their own particular ability levels-tests that may be very different for each student. (Readers interested in further information on computer-adaptive language testing should see Larson & Madsen, 1985, and for a discussion of both positive and negative aspects, see Canale, 1986.)

One example of the actual development of a CALT is the Montgomery County Public Schools project which is described in Stevenson and Gross (1991). Madsen (1991) describes another example of a CALT which was applied to students at Brigham Young University in Utah for testing reading and listening abilities. The Madsen (1991) study indicates that many fewer items are necessary in administering computer-adaptive language tests than are necessary in pencil-and-paper tests and that the testing time is correspondingly shorter. For example, the CALT in Madsen (1991) used an average of 22.8 items to adequately test the students in an average of 27.2 minutes. The comparable conventional reading test used in the study required 60 items and 40 minutes.

Educational Testing Service (ETS) is providing considerable leadership in the area of what they are calling computer-based tests. That organization is already offering the GRE and PRAXIS as computer-based tests in 180 countries. In 1998, a computer-based version of the TOEFL examination will be released in North America and selected countries abroad, though paper-and-pencil versions will continue to be used until computer delivery is available.

Because of recent efforts to develop computer-based versions of the TOEFL, some of the research at ETS has focused on the effects of computer familiarity on TOEFL test performance. For instance, Kirsch, Jamieson, Taylor, and Eignor (1997) (based on a computer-familiarity scale discussed in Eignor, Taylor, Kirsch, & Jamieson, 1997) indicates that, in general, computer familiarity is related to individuals' TOEFL scores. At the same time, Taylor, Jamieson, Eignor, and Kirsch (1997) indicates that, after students participate in a computer-based testing tutorial, there is no meaningful relationship between

computer familiarity and individuals' TOEFL scores. (Readers interested in more information on the new TOEFL developments can contact TOEFL Programs and Services, P.O. Box 6155, Princeton, NJ 08541-6155, use e-mail: toefl@ets.org, or visit their web site: http://www.toefl.org. Readers interested in more details about computer-adaptive testing in general would benefit from reading the educational measurement "primer" on the topic: Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg, & Thissen, 1990. For a fairly technical book on the subject, see Weiss, 1983.)

Effectiveness of Computers in Language Testing

Educational Testing Service (1996) claims the following advantages for their new computer-based TOEFL:

1. Further enhancements to test design before 2000
2. Greater flexibility in scheduling test administrations
3. Greater standardization of test administration conditions
4. Portions of test individualized to examinee ability level
5. Inclusion of writing with every test administration
6. Examinee choice of handwriting or typing essay
7. Ability to record multiple aspects of examinee test-taking behavior
8. Platform for future innovations in test design and services (p. 5).

Judging by what they are claiming, at least portions of this new computer-based test will eventually be computer-adaptive.

Brown (1992b) looked in more detail at both the advantages and disadvantages of using computers in language testing. That discussion will be expanded next.

Advantages. The advantages of using computers in language testing can be further subdivided into two categories: testing considerations and human considerations.

Among the *testing considerations*, the following are some of the advantages of using computers in language testing:

1. Computers are much more accurate at scoring selected-response tests than human beings are.
2. Computers are more accurate at reporting scores.
3. Computers can give immediate feedback in the form of a report of test scores, complete with a printout of basic testing statistics.
4. IRT and computer-adaptive testing allow testers to target the specific ability levels of individual students and can therefore provide more precise estimates of those abilities (see Bock & Mislevy, 1982).
5. The use of different tests for each student should minimize any practice effects, studying for the test, and cheating (for discussion of an IRT strategy to help spot such cheating, see Drasgow, Levine, & McLaughlin, 1987).
6. Diagnostic feedback can be provided very quickly to each student on those items answered incorrectly if that is the purpose of the test. Such feedback can even be fairly descriptive if artificial intelligence is used (for more on such uses of artificial intelligence, see Baker, 1989, pp. 423-425, or Bunderson, Inouye, & Olsen, 1989, pp. 398-402).

Among the *human considerations*, the following are some advantages of using computers in language testing:

1. The use of computers allows students to work at their own pace.
2. CALTs generally take less time to finish than traditional paper-and-pencil tests and are therefore more efficient (as found for CALTs in Madsen, 1991, and for CATs in Kaya-Carton, Carton, & Dandonoli, 1991, and Laurier, 1996).
3. In CALTs, students should experience less frustration than on paper-and-pencil tests because they will be working on test items that are appropriate for their own ability levels.
4. Students may find that CALTs are less overwhelming (as compared to equivalent paper-and-pencil tests) because the questions are presented one at a time on the screen rather than in an intimidating test booklet with hundreds of test items.
5. Many students like computers and even enjoy the testing process (Stevenson & Gross, 1991).

Disadvantages. The disadvantages of using computers in language testing can also be further subdivided into two categories: physical considerations and performance considerations.

Among the *physical considerations*, the following are some of the disadvantages of using computers in language testing:

1. Computer equipment may not always be available, or in working order. Reliable sources of electricity are not universally available.
2. Screen capacity is another physical consideration. While most computers today have overcome the 80 characters by 25 lines restrictions of a few years ago, the amount of material that can be presented on a computer screen is still limited. Such screen size limitations could be a problem, for example, for a group of teachers who wanted to develop a reading test based on relatively long passages.
3. In addition, the graphics capabilities of many computers (especially older ones) may be limited, and even those machines that do have graphics may be slow (especially the cheaper machines). Thus, tests involving even basic graphs or animation may not be feasible at the moment in many language teaching situations.

Among the *performance considerations*, the following are some of the disadvantages of using computers in language testing:

1. The presentation of a test on a computer may lead to different results from those that would be obtained if the same test were administered in a paper-and-pencil format (Henning, 1991). Some limited research indicates that there is little difference for math or verbal items presented on computer as compared with pencil-and-paper version (Green, 1988) or on a medical technology examination (Lunz & Bergstrom, 1994), but much more research needs to be done on various types of language tests and items.
2. Differences in the degree to which students are familiar with using computers or typewriter keyboards may lead to discrepancies in their performances on computer-assisted or computer-adaptive tests (Hicks, 1989; Henning, 1991; Kirsch, Jamieson, Taylor, & Eignor, 1997)
3. Computer anxiety (i.e., the potential debilitating effects of computer anxiety on test performance) is another potential disadvantage (Henning, 1991).

What Issues Have Already Received Attention?

Judging by the first half of this paper, language testers know a great deal about computers in language testing. For instance, we know:

1. How to use an item bank to create, pilot, analyze (with item response theory), store, manage, and select test items for purposes of making multiple test forms.
2. How to create more imaginative computer-assisted language tests that take advantage of the new technologies.
3. How to build computer-adaptive tests that will (a) help us select and fit items to individual students' abilities, (b) help us know when to end the test when the student's ability level is found, and hence, (c) help us build computer-adaptive tests that are relatively short in terms of number of items and time.
4. How effective computers are in language testing, including the advantages (in terms of testing and human considerations) and disadvantages (in terms of physical and performance considerations).

At the same time, language testers have a great deal to learn about computers in language testing, as I will explain next.

FUTURE DIRECTIONS: COMPUTER-ADAPTIVE LANGUAGE TESTING

Most of the research cited so far has been practical in nature. In my view, research on computers in language testing will inevitably become more technical, complex, and detailed. Drawing on the wider literature of the educational testing field, I found that the educational measurement literature on using computers in testing developed very much like the language testing literature has (for an overview of these developments, see Bunderson, Inouye, & Olsen, 1989). However at this point in time, researchers in educational measurement have developed well beyond the practical questions of how to item bank, how to definitions and distinguish between computer-assisted and computer-adaptive testing, and how to measure the effectiveness of computers in testing. In short, they are now addressing considerably more technical, complex, and detailed issues.

Examining the types of concerns that have arisen recently for educational testers may therefore point the way for future research on computers in language testing and provide a basis for researchers in our field who would like to begin studying these issues. Because of length constraints, I will narrow my exploration to recent concerns about computer-adaptive testing in order to illustrate how at least one strand of research points to the future for language testers. The discussion will be organized into three categories--CALT design issues, CALT scoring issues, and CALT logistical issues--under subheadings that pose the central questions that seem to be the focus of research into computer use in educational testing. All of these questions will be directed at computer-adaptive tests, as are the studies in the educational testing literature.

CALT Design Issues

A number of CALT design issues need to be addressed by language testers before we can be fairly sure we are using such tests responsibly, including at least the following questions: How should we pilot CALTS? Should a CALT be standard length or vary across students? How should we sample CALT items? What are the effects of changing the difficulty of CALT items? How can we deal with item sets on a CALT?

How should we pilot CALTs? The problem that must be addressed here is caused by the fact that CALTs, because of their adaptive nature, may be different for every student. In classical theory tests, the examinations were the same for every student. Hence, a single examination could be piloted, item analyzed, revised, and then validated during the operational administration. In contrast, a new CALT is

created each time a student takes the test (for a quick overview of this process, see Baker, 1989, pp. 425-426).

In educational testing circles, the strategy that is used to overcome the piloting problem is called simulation. Using item response theory, testers can simulate the answers of typical students based on the pilot responses of real students. By simulating such responses in various combinations and for different test lengths, researchers can study the probable characteristics of the test under many different conditions and lengths (as shown in Eignor, Way, Stocking, & Steffen, 1993). To my knowledge, no such simulation studies have been done in the language testing context.

Should a CALT be standard length or vary across students? One of the advantages of computer-adaptive testing is that the test can be tailored to each student's ability levels. Such tailoring results in a test that may be more efficient at some levels of ability, requiring fewer items to reliably estimate the students' abilities, but less efficient at other levels, requiring more items to do a reliable job. The issue that arises then is whether the test should be the same length for each student or of different lengths tailored to each student's ability.

Stocking (1987) indicates that the tailored strategy resulting in different lengths may produce biased scores for students who end up taking short versions. If the test is made to be relatively long for all students, even for those students who will be taking the "short" versions, no such biases may surface. However, if the test is short for all students, it may be preferable to use the same length test for all students. These important length issues have yet to be investigated for CALTs. Perhaps they should be.

How should we sample CALT items? Traditional tests, in order to be theoretically valid, are often designed on the basis of clearly developed test specifications to sample different knowledges or skills that make up a particular domain. Because CALTs will typically be shorter than traditional tests, testing all of the specifications may be impossible. Items could be randomly selected from all the possible specifications. However, if CALTs are short for only some students, a better strategy might be to develop a sampling algorithm that takes into account the issue of specification sampling by keeping track of which specifications have been sampled for a given student and selecting the remaining items in a way that best fulfills the rest of the test specifications. Such a scheme might even take into account the relative importance of the various specifications. Naturally, any such scheme should be based on research-research that has yet to be conducted in relation to CALTs (for related research in CAT, see Stocking & Swanson, 1993; Swanson & Stocking, 1993).

What are the effects of changing the difficulty of CALT items? Bergstrom, Lunz, and Gershon (1992) found that altering the difficulty of CAT items slightly raises the number of items that are necessary to test students with adequate precision, but otherwise has little affect on the estimation of their abilities. They did so by studying the responses of 225 students on a medical technology examination. No such study has been conducted on language tests to my knowledge. Is there a relationship between item difficulty and test length on a CALT, but little effect on ability estimation in language testing? And if both sets of results are the same in language testing as in the education literature, what difficulty level would be ideal in terms of language test length and precision of measurement in relation to the level of language proficiency of the students?

How can we deal with item sets on a CALT? A problem that occurs whenever item banking is done, even for paper-and-pencil tests, is how to deal with item sets (sets of items determined by reading or listening passages, or sets of items based on item formats, topics, item contents, etc.) Good reasons may exist for keeping certain items together and even for ordering them in a certain way. Wainer and Kiely (1987) call these sets of items "testlets." (Wainer et al., 1990, also discuss testlets, as did Sheehan & Lewis, 1992,

and Bunderson, Inouye, & Olsen, 1989, p. 393-394, 398, refer to them as "reference tasks," while Boekkooi-Timminga, 1990, explores a related concept she calls item clustering.)

Unfortunately, using item sets, or testlets, may result in a selection of test items that are not maximally efficient from an IRT perspective. Since such item sets often occur in language tests, especially in academic reading and lecture listening tests, the issue of item sets is an important one in language testing. Hence, investigations should be conducted into the best strategies to use in dealing with CALT item sets, as well as on the relative effectiveness of item sets as compared to independent items. Research might also be profitable on the ordering of items within sets, and indeed, the ordering of items throughout a CALT.

CALT Scoring Issues

A number of CALT scoring issues also need to be addressed by language testers before we can be fairly sure we are using such tests responsibly, including at least the following questions: How should we score CALTS? How should we deal with CALT item omissions? How should we make decisions about cut-points on CALTs?

How should we score CALTs? A number of alternative methods have been explored in the educational literature including at least the following:

1. Raw scores (a simple sum of the number of correct items)
2. Weighted raw scores (in which some items count more than others)
3. Scaled scores (e.g., the TOEFL, which is equated across forms; for an explanation of scaling and equating CATS, see Dorans, 1990; for studies using the techniques, see Mazzeo, Druesne, Raffeld, Checketts, & Muhlstein, 1991; O'Neill, Folk, & Li, 1993; Schaeffer, Steffen, & Golub-Smith, 1993)
4. Any of the above (1-3) corrected for guessing (as discussed for CATs in Stocking, 1994)
5. Any of the above (1-3) based on polytomous (as opposed to dichotomous) scoring (for a study of this issue on CATs, see Dodd, De Ayala, & Koch, 1995)
6. Any of the above (1-4) scores referenced to a conventional test (as described for CATs in Ward, 1988; Wainer et al., 1990), especially when a CALT is first introduced to replace an existing paper-and-pencil test
7. Any of the above (1-4) scores referenced to a set of anchor items (anchor items are typically a set of items that all students take along with the rest of the test).

Research on which method works best for which purposes when interpreting CALT scores would be very beneficial. Also, when using IRT for the development of CALTs, language testers must grapple with explaining the scoring methods in non-technical terms so that students and teachers can understand the results. Hence, CALT research might profitably focus on how to best convey such IRT scoring information to lay people.

How should we deal with CALT item omissions? On traditional tests, especially those scored with a correction for guessing, students can and do omit items if they are not able to answer them. However, on a CALT, problems arise if the students omit one or more items. How should testers deal with such omissions? Wainer et al. (1990, pp. 236-237) chose to not allow for omissions on their CATs; the students simply could not go on to the next item until they had answered the one in front of them.Another strategy would be to allow omissions but assume that omitted items are wrong. If that strategy was followed, what would be the effect of such a wrong item on the estimation of the items that should follow?

Other problems may arise if students are allowed to omit items. For instance, if omitted items are not scored, students can manipulate the test by skipping items until they find items they can answer correctly. According to Lunz and Bergstrom (1994), such students would receive undeservedly high scores. Another problem is that students can simply omit all items and get a good look at all of the items in the item bank. This would have important implications for item exposure (as discussed below). Another possibility is that students at different language proficiency levels might omit items in different ways. Hence, omission patterns would be linked to language proficiency as measured on such a test and would become a source of measurement error. All of these issues related to item omissions on CALTs need to be researched and resolved in one way or another.

How should we make decisions about cut-points on CALTs? The literature on deciding cut-points is often referred to as standards setting (for a review, see Jaeger, 1989, pp. 491-500). Reckase (1983) has suggested the use of the sequential probability ratio test (SPRT) for making pass-fail decisions in adaptive and other related tests. The SPRT was originally developed by Wald (1947) as a quality control device, but both Reckase (1983) and Kingsbury and Weiss (1983) show how the SPRT can be applied to adaptive tests as well. Lewis and Sheehan (1990) suggested using Bayesian decision theory for making mastery/non-mastery decisions on a computerized sequential mastery test. Du, Lewis, and Pashley (1993) demonstrated the use of a fuzzy set approach for the same purpose. The SPRT and other methods just described should all be considered in terms of how well they help with test constuction and decision making on CALTs.

CALT Logistical Issues

In addition to the important logistical issues raised by Green (1990), which included system, hardware, human, and software issues, at least three CALT logistical issues need to be addressed by language testers: How can we avoid CALT item exposure? Should we provide item review opportunities on a CALT? How can we comply with legal disclosure laws when administering CALTs?

How can we avoid CALT item exposure? Item exposure occurs when students see any given item. Since exposure means that other students who might take the test in the future may know about the exact content of those items which have been exposed, such items should not be used again. In traditional tests, large numbers of students are tested on one day with a relatively small number of items, all in one exposure. Thus, even though the items could not be used again, they were used rather efficiently to test a large number of students. However, on a CALT, even though each student gets a different test form, unless there is an infinite number of items, whatever items are in the item bank will be exposed rather quickly. If the test is administered on a daily walk-in basis at a computer lab, for instance, all items in the bank could be exposed within a matter of days without having tested very many students. In addition, as discussed above, particularly in a situation were item omissions are permitted on a CALT, students can simply omit all items and get a good look at all of the items in the item bank. Naturally, such wholesale item exposure would have important implications for test security. The steps that can be taken to slow down the process of item exposure are to:

1. Have a very large item bank with a wide variety of difficulty levels to meet all item specifications desired in the test
2. Have the computer select a number of items which might come next (rather than a single item) and then randomly select from among those possibilities (as in McBride & Martin, 1983).
3. Have the computer select the next item based on complex probabilistic models like those discussed in Stocking (1992), Stocking and Lewis (1995a; 1995b).
4. Use simulation studies to estimate the efficiency of different sized item banks in minimizing exposure, then stay within whatever size limits those studies suggest.

5. Circulate item banks, or sub-banks, through different testing sites.
6. Constantly change the items in an item bank by adding new items and eliminating old ones (especially those most likely to have been exposed).
7. Monitor the functioning of items within the item pool by keeping track of students' item performances, and identify items that appear to have been exposed

Clearly, studies should be conducted into the relative effectiveness of strategies for dealing with CALT item exposure.

Should we provide item review opportunities on a CALT? In traditional paper-and-pencil testing situations, students who have time remaining at the end of the test can go back and review earlier items. Given the fact that CALT algorithms select the next item based on previous item performances, allowing students to go back and review items would undermine the theoretical nature of the test. Lunz, Bergstrom, and Wright (1992) indicate that this is not a big problem for CATs, but what about CALTs?

Wainer (1992) suggests that very testwise students could use reviewing as a way of manipulating the test and gaining an unfair edge. The degree to which these item reviewing issues are important to CALTs is yet to be determined. (For a fuller explanation of testing algorithms in CATs, see Thissen & Mislevy, 1990.)

How can we comply with legal disclosure laws when administering CALTs? New York state has so-called "truth in testing" disclosure laws dating back to 1979 that require that standardized tests administered in the state be made available to examinees for inspection within 30 days of the administration. Naturally, with today's communications technology, if a test must be disclosed in one state, it might as well be disclosed in all states and around the world. That requirement does not cause undue problems for paper-and-pencil tests which can be administered to tens or even hundreds of thousands of students in a one month period. However, a problem arises from the fact that a relatively large item bank is necessary for CALTs. To disclose the entire item bank of a CALT every 30 days, or even to disclose to students only those items they took, would be very costly in terms of development time and manpower to produce, pilot, analyze, and distribute new items. Currently, the New York legislature is considering laws (SO3292-C or A5648-C) that will continue the basic requirements of the 1979 law, but be updated to more reasonably fit with the new CATs and the logistics involved (see Lissitz, 1997). However, in the interim, research should be conducted into the best strategies for developing new item banks and for phasing out old ones and disclosing them publicly.

CONCLUSION

The purpose of this paper is to examine recent developments in language testing that directly involve computers. In the process, I have looked at what language testers have learned in the specific area of CALT and found substantial information on:

1. How to use an item bank
2. How to use new technologies
3. How to build computer-adaptive tests
4. How effective computers are in language testing.

Next, I examined the educational testing literature for ideas on what directions future research on computers in language testing might take, focusing on the dominant issue of computer-adaptive testing and found that future research might benefit from answering the following questions:

1. How should we pilot CALTS?
2. Should a CALT be standard length or vary across students?
3. How should we sample CALT items?
4. What are the effects of changing the difficulty of CALT items?
5. How can we deal with item sets on a CALT?
6. How should we score CALTS?
7. How should we deal with CALT item omissions?
8. How should we make decisions about cut-points on CALTs?
9. How can we avoid CALT item exposure?
10. Should we provide item review opportunities on a CALT?
11. How can we comply with legal disclosure laws when administering CALTs?

However, I would like to stress that the computer-adaptive testing involved in the above 11 questions is only one stream of computer-related research in education, psychology, and related fields. Important development and research are also going on in areas like: (a) testing in intelligent teaching systems, (b) testing using the Internet, (c) handwriting and speech recognition, (d) analysis and scoring of open-ended responses (like compositions and speech samples), and (e) alternative psychometric models for analyzing the results of the more complex information that can be gathered using computer-assisted response analysis. (Readers interested in exploring these issues further might start by reading Alderson, 1991; Bejar & Braun, 1994; Burstein, Frase, Ginther, & Grant, 1996; Corbel, 1993; Jamieson, Campbell, Norfleet, & Berbisada, 1993; Mislevy, 1993, 1994; Powers, Fowles, Farnum, & Ramsey, 1994.)

Naturally, through all of our computer-related research efforts, professional quality language testing should continue to be our goal, and such testing should continue to adhere to the *Standards for educational and psychological testing* (American Psychological Association, 1985) agreed to by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education for all educational and psychological tests. Special guidelines have also been published (American Psychological Association, 1986), which interpret the above *Standards* in terms of how they should be applied to computer-based testing and score interpretations (see also Green, Bock, Humphreys, Linn, & Reckase, 1984). In keeping with those two sets of guidelines, ongoing research must also be conducted on the reliability and validity of each and every CALT that is developed in the future.[1]

Notes 1 The reliability issues for CATs are addressed in Thissen, 1990; the validity issues are addressed in Steinberg, Thissen, & Wainer, 1990; and reliability and validity were studied together in McBride & Martin, 1983, and Bennett & Rock, 1995.

ABOUT THE AUTHOR

James Dean ("JD") Brown, Professor on the graduate faculty of the Department of ESL at the University of Hawai'i, has published numerous articles on language testing and curriculum development. He has authored books on language research methods (Cambridge), curriculum development (Heinle & Heinle), and testing (Prentice-Hall), and edited a book (with Yamashita) on language testing in Japan (JALT).

E-mail: brownj@hawaii.edu

REFERENCES

Alderson, C. J. (1991). Innovation in language testing: Can the microcomputer help? *Language Testing Update*, Special Report No. 1.

American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Psychological Association. (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: American Psychological Association.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Baker, F. B. (1989). Computer technology in test construction and processing. In R. L. Linn *Educational measurement* (3rd ed., pp. 409-428). London: Collier Macmillan.

Bejar, I., & Braun, H. (1994). On the synergy between assessment and instruction: Early lessons from computer-based simulations. *Machine-Mediated Learning, 4*, 5-25.

Bennett, R. E., & Rock, D. A. (1995). Generalizability, validity, and examinee perceptions of a computer-delivered formulating-hypotheses test. *Journal of Educational Measurement, 32*, 19-36.

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the difficulty level in computer adaptive tests. *Applied Measurement in Education, 5*, 137-149.

BestTest [Computer software]. (1990). Chicago, IL: WiseWare.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Boekkooi-Timminga, E. (1990). A cluster-based method for test construction. *Applied Psychological Measurement, 14*, 341-354.

Brown, J. D. (1992a). Technology and language education in the twenty-first century: Media, message, and method. *Language Laboratory, 29*, 1-22.

Brown, J. D. (1992b). Using computers in language testing. *Cross Currents, 19*, 92-99.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 367-407). London: Collier Macmillan.

Burstein, J., Frase, L., Ginther, A., & Grant, L. (1996). Technologies for language assessment. *Annual Review of Applied Linguistics, 16*, 240-260.

Canale, M. (1986). The promise and threat of computerized adaptive assessment of reading comprehension. In C.W. Stansfield (Ed.), *Technology and language testing* (pp. 29-45). Washington, DC: TESOL.

Corbel, C. (1993). Computer-enhanced language assessment. In G. Brindley (Ed.), *Research Report Series 2.* Sydney: National Centre for English Language Teaching and Research, Marquarie University.

Corel Paradox 7.0 [Computer software]. (1996). Ottawa, Ontario, Canada: Corel Corporation

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*, 5-22.

Dorans, N. J. (1990). Scaling and equating. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 137-160). Hillsdale, NJ: Lawrence Erlbaum.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59-80.

Du, Y., Lewis, C., & Pashley, P. J. (1993). Computerized mastery testing using fuzzy set decision theory. *Applied Measurement in Education, 6*, 181-193.

Dunkel, P. (1991). The effectiveness research on computer-assisted instruction and computer assisted language learning. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 5-36). New York: Newbury House.

Educational Testing Service. (1996). *TOEFL: Announcing computer-based testing.* Princeton, NJ: Educational Testing Service.

Eignor, D., Taylor, C., Kirsch, I., & Jamieson, J. (1997). Development of a scale for assessing the level of computer familiarity of TOEFL examinees. Unpublished ms. Princeton, NJ: Educational Testing Service.

Eignor, D. R., Way, W. D., Stocking, M. L., & Steffen, M. (1993). *Case studies in computer adaptive test design through simulation* (Research report # 93-56). Princeton, NJ: Educational Testing Service.

Flaugher, R. (1990). Item pools. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 41-63). Hillsdale, NJ: Lawrence Erlbaum.

Green, B. F. (1988). Construct validity of computer-based tests. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 77-86). Hillsdale, NJ: Lawrence Erlbaum.

Green, B. F. (1990). System design and operations. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 23-40). Hillsdale, NJ: Lawrence Erlbaum.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. B., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-360.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Item response theory.* Newbury Park, CA: Sage.

Henning, G. (1986). Item banking via dBase II: The UCLA ESL Proficiency Examination experience. In C. W. Stansfield (Ed.), *Technology and language testing* (pp. 69-77). Washington, DC: TESOL.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research.* New York: Newbury House.

Henning, G. (1991). Validating an item bank in a computer-assisted or computer-adaptive test. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 209-222). New York: Newbury House.

Henning, G., Johnson, P. J., Boutin, A. J., & Rice, H. R. (1994). Automated assembly of pre-equated language proficiency tests. *Language Testing, 11*, 14-28.

Hicks, M. (1989). *The TOEFL computerized placement test: Adaptive conventional measurement.* TOEFL Research Report No. 31. Princeton, NJ: Educational Testing Service.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). London: Collier Macmillan.

Jamieson, J., Campbell, J., Norfleet, L., & Berbisada, N. (1993). Reliability of a computerized scoring routine for an open-ended task. *System, 21*, 305-322.

de Jong, J. H. A. L. (1986). Item selection from pretests in mixed ability groups. In C. W. Stansfield (Ed.), *Technology and language testing* (pp. 91-108). Washington, DC: TESOL.

Kaya-Carton, E., Carton, A. S., & Dandonoli, P. (1991). Developing a computer-adaptive test of French reading proficiency. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 259-284). New York: Newbury House.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257-283). New York: Academic Press.

Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1997). Computer familiarity among TOEFL examinees. Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Larson, J. W., & Madsen, H. S. (1985). Computerized adaptive language testing: Moving beyond computer-assisted testing. *CALICO Journal, 2*, 32-36, 43.

Laurier, M. (1991). What we can do with computerized adaptive testing...and what we cannot do! In S. Anivan (Ed.), *Current developments in language testing* (pp. 244-255). Singapore: Regional Language Centre.

Laurier, M. (1996). Using the information curve to assess language CAT efficiency. In A. Cumming & R. Berwick (Eds.),*Validation in language testing* (pp. 111-123). Clevedon, UK: Multilingual Matters.

Lewis, C., Sheehan, K. (1990). Using Bayesian decision theory to design a computerized adaptive mastery test. *Applied Psychological Measurement, 14*, 367-386.

Lissitz, B. (1997). The New York standardized testing legislation. (p. 2). *National Council on Measurement in Education Quarterly Newsletter (5)*2.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement, 31,* 251-263.

Lunz, M. E., & Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computer adaptive tests. *Applied Psychological Measurement, 16*, 33-40.

Madsen, H. S. (1991). Computer-adaptive testing of listening and reading comprehension. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 237-257). New York: Newbury House.

Madsen, H. S., & Larson, J. W. (1986). Computerized Rasch analysis of item bias in ESL tests. In C. W. Stansfield (Ed.), *Technology and language testing* (pp. 47-67). Washington, DC: TESOL.

Mazzeo, J., Druesne, B., Raffeld, P. C., Checketts, K. T., & Muhlstein, A. (1991). *Comparability of computer and paper-and-pencil scores for two CLEP general examinations* (Report #91-5). New York: College Entrance Examination Board.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223-236). New York: Academic Press.

MicroCAT Testing System [Computer software]. (1984). St. Paul, MN: Assessment Systems.

Microsoft Access 2.0 [Computer software]. (1996). Redmond, WA: Microsoft Corporation.

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19-39). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483.

Mislevy, R. J., & Bock, R. D. (1986). *PC-BILOG: Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software.

Neu, J., & Scarcella, R. (1991). Word processing in the ESL writing classroom: A survey of student attitudes. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 169-187). New York: Newbury House.

O'Neill, K., Folk, V., & Li, M.-Y. (1993). *Report on the pretest calibration study for the computer-based academic skills assessments of The Praxis Series: Professional assessments for beginning teachers.* Princeton, NJ: Educational Testing Service.

PARGrade 3.0 [Computer software]. (1990). Costa Mesa, CA: Economics Research.

PARScore 3.0 [Computer software]. (1990). Costa Mesa, CA: Economics Research.

PARTest 3.0 [Computer software]. (1990). Costa Mesa, CA: Economics Research.

Phinney, M. (1991). Computer-assisted writing and writing apprehension in ESL students. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 189-204). New York: Newbury House.

Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1994). Will they think less of handwritten essays if others wordprocess theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement, 31,* 220-233.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237-255). New York: Academic Press.

Reid, J. (1986). Using the writer's workbench in composition teaching and testing. In C.W. Stansfield (Ed.), *Technology and language testing* (pp. 167-188). Washington, DC: TESOL.

Schaeffer, G., Steffen, M., & Golub-Smith, M. (1993). *Introduction of a computer adaptive GRE General Test* (Research Report # 93-57). Princeton, NJ: Educational Testing Service.

Sheehan, K., & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement, 16,* 65-76.

Steinberg, L., Thissesn, D., & Wainer, H. (1990). Validity. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 187-231). Hillsdale, NJ: Lawrence Erlbaum.

Stenson, H. (1988). *Testat: A supplementary module for SYSTAT* (version 2.0). Chicago, IL: SYSTAT.

Stevenson, J., & Gross, S. (1991). Use of a computerized adaptive testing model for ESOL/bilingual entry/exit decision making. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 223-235). New York: Newbury House.

Stocking, M. L. (1987). To simulated feasibility studies in computerized adaptive testing. *Applied Psychology: An International Review, 35,* 263-277.

Stocking, M. L. (1992). *Controlling item exposure rates in a realistic adaptive testing paradigm* (Research Report # 93-2). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1994). *An alternative method for scoring adaptive tests* (Research Report # 94-48). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1995a). *Controlling item exposure conditional on ability in computerized adaptive tests* (Research Report #95-24). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1995b). *A new method of controlling item exposure in computerized adaptive tests* (Research Report #95-25). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277-292.

Suen, H. K. (1990). *Principles of test theories.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Swanson, L. & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17,* 151-166.

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1997). Measuring the effects of computer familiarity on computer-based language tasks. Unpublished manuscript. Princeton, NJ: Educational Testing service.

Testmaster [Computer software]. (1988). Zurich, Switzerland: Eurocentres.

Thissen, D. (1990). Reliability and measurement precision. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A prime*r (pp. 161-186). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., & Mislevy, R. J. (1990). Testing algorithms. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 103-134). Hillsdale, NJ: Lawrence Erlbaum.

Tung, P. (1986). Computerized adaptive testing: Implications for language test developers. In C.W. Stansfield (Ed.), *Technology and language testing* (pp. 11-28). Washington, DC: TESOL.

Wainer, H. (1992). *Some practical considerations when converting a linearly administered test to an adaptive format* (Research Report #92-13). Princeton, NJ: Educational Testing Service.

Wainer, H., Dorans, N. J., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). Future challenges. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A prime*r (pp. 233-272). Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H. C., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-201.

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration and proficiency estimation. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 65-102). Hillsdale, NJ: Lawrence Erlbaum.

Wald, A. (1947). *Sequential analysis.* New York: Wiley.

Ward, W. C. (1988). The College Board Computerized Placement Tests: An application of computerized adaptive testing. *Machine-Mediated Learning, 2,* 217-282.

Weiss, D. J. (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Wright, B. D., Linacre, J. M., & Schulz, M. (1990). *BIGSTEPS: General-purpose Rasch analysis program* (version 2.00). Chicago, IL: Mesa Press.s